



TALEND DATA PREPARATION FREE DESKTOP

GUIDE DE PRISE EN MAIN V2.5



GUIDE DE PRISE EN MAIN DE TALEND DATA PREPARATION

Pour vous rendre à une section spécifique du guide, cliquez sur l'une des sections suivantes.

- 01 PRÉSENTATION DE TALEND DATA PREPARATION
- 02 ACCÈS ET DÉMARRAGE DE TALEND DATA PREPARATION
- 03 EXERCICES FACILES DE NETTOYAGE DE DONNÉES
- 04 OPÉRATIONS BASIQUES DE MANIPULATION DE DONNÉES
- 05 NETTOYAGE ET FORMATAGE DE DATES

GUIDE DE PRISE EN MAIN DE TALEND DATA PREPARATION

Présentation de Talend
Data Preparation

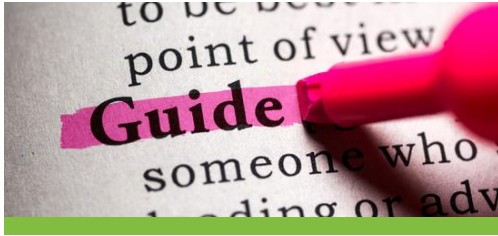
Accès et démarrage de
Talend Data Preparation

Exercices faciles de
nettoyage de données

Opérations basiques de
manipulation de données

Nettoyage et formatage
de dates

À propos de ce guide



Le guide de prise en main de Talend Data Preparation fournit des instructions étape par étape qui vous permettront de créer et de réaliser de A à Z des scripts de préparation de données.



La démo été conçue à partir de cas d'utilisation réels avec des données marketing. Ils reflètent les problématiques auxquelles de nombreux utilisateurs font face tous les jours, sur des tableurs les obligeant à utiliser des macros complexes, voire des scripts VBA.



Le but est de vous familiariser avec cet outil aussi intuitif qu'une interface utilisateur sur le web. Vous comprendrez comment Talend peut vous aider à découvrir, nettoyer, mettre en forme et enrichir vos données.

À propos de Talend Data Preparation



Des données nettoyées et utiles en seulement quelques minutes (au lieu de plusieurs heures)

- Un point d'entrée unique pour tout type de source de données
- Découverte, nettoyage et formatage interactifs
- Automatisez et réutilisez les formules de préparation de données



L'outil de nettoyage de données en libre-service pour tous

- Mettez les données à votre service dans vos tâches quotidiennes
- Découvrez et explorez vos données
- Laissez vous guider sur votre chemin vers des données actionnables



Donnez de l'autonomie aux tâches informatiques et profitez plus rapidement de connaissances approfondies

- Accès aux données maîtrisé et en libre-service pour tous
- Évitez des incohérences et des fuites de données nuisibles à l'entreprise
- Soulagez les équipes informatiques et dynamisez la productivité



**Regardez la vidéo de présentation
de Talend Data Preparation**

Si vous avez déjà installé Talend Data Preparation,
cliquez ici pour passer les instructions d'installation.

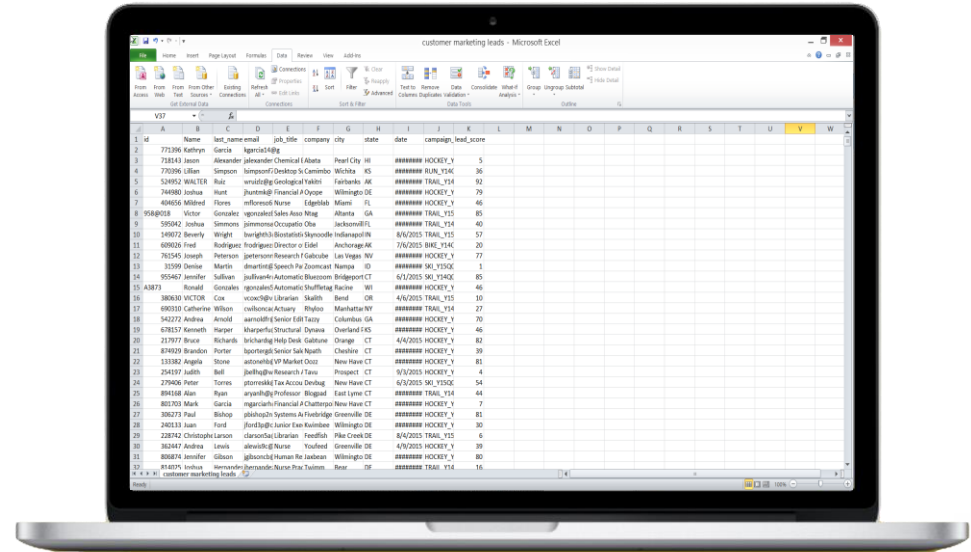
Marketing Lead Data Preparation

Les données du fichier “customer marketing leads.csv” concernent des leads. Elles comportent des problèmes de qualité et de nombreux champs doivent donc être reformatés. Analyser les données brutes dans ce fichier aboutirait à des résultats décevants, à cause d’informations incorrectes ou de valeurs manquantes dont la correction avec Excel prendrait des heures.

Dans cette démo, nous vous accompagnerons dans des actions faciles de préparation de données qui, sur Excel, vous donneraient du fil à retordre.

Vous découvrirez comment :

- Changer rapidement les valeurs des données après les avoir identifiées grâce à des graphiques et à des filtres très simples, sans codage !
- Tester des fonctionnalités telles que les histogrammes pour corriger les données !
- Manipuler du texte, des dates et des champs numériques dans le même fichier en seulement quelques clics !



Prérequis pour Talend Data Preparation

Prérequis matériel

Processeur	Processeur 64-bit requis
Mémoire allouée	Minimum 1GB
Espace sur le disque	Minimum 500MB + datasets = 5GB recommandés

Prérequis logiciels

Système d'exploitation	<ul style="list-style-type: none">Windows 7 64-bit ou plus récentMac OS X 10.7 "Lion" ou plus récent
------------------------	---

Navigateurs Web compatibles

Mozilla Firefox / Firefox ESR	Dernière version
Microsoft Internet Explorer	11
Microsoft Edge	Dernière version
Apple Safari	10
Google Chrome	Dernière version

Voici les informations à propos des logiciels et du matériel recommandés pour commencer avec Talend Data Preparation.

Java:

Il n'y a pas de besoin spécifique pour la plupart des ordinateurs Windows et Apple. Cependant, si vous souhaitez installer la version Apache de Talend Data Preparation, vous devez avoir Oracle Java 8 (64 bits) installé sur votre ordinateur. La version par défaut pour Windows 32 bits n'est pas supportée, seule la version 64 bits est supportée.

Comment télécharger Talend Data Preparation?



Téléchargez Talend Data
Preparation ici.

Choisissez votre
système d'exploitation
et le téléchargement
démarré
automatiquement.

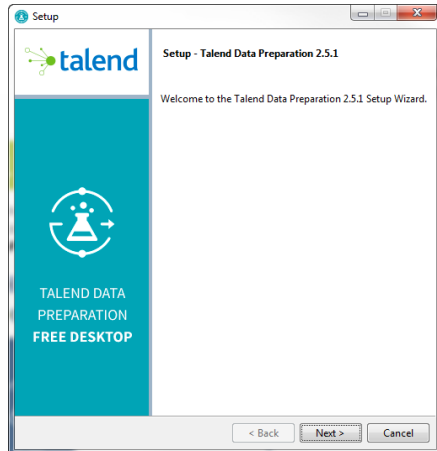
Comment installer Talend Data Preparation sur Windows ?



La version pour Windows est fournie en tant qu'installateur Microsoft Windows standard. Vous aurez besoin des droits administrateur pour l'exécuter. Pour installer et démarrer Talend Data Preparation veuillez suivre les étape suivantes :

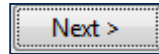
1

Après avoir téléchargé le fichier, double-cliquez sur **Talend-DataPreparation-Free-Desktop-2.5.exe**



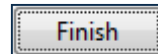
2

Cliquez sur **Next** pendant l'installation et utilisez les paramètres par défaut.



3

Cliquez sur **Finish** une fois l'installation terminée.



4

Pour commencer à utiliser Talend Data Preparation, cliquez sur l'icône du programme ou sur le raccourci dans le menu Démarrer.



Installation alternative pour les utilisateurs Windows



Si vous n'avez pas les droits administrateurs requis pour utiliser l'installateur, suivez les étapes ci-dessous pour installer Talend Data Preparation via un fichier .zip :

1

Sur la page de téléchargement de Talend Data Preparation, descendez jusqu'à la section **Versions and Releases Notes** et clique sur **Other Releases**.

2

Téléchargez le fichier **Talend-DataPreparation-Free-Desktop-windows-2.5.1.zip**.

3

Dézippez le fichier où vous le souhaitez sur votre ordinateur.

4

Lancez le fichier **.exe** pour utiliser l'outil Talend Data Preparation.

Other Releases

File Name	Version	Release Date	Release Type	Supported Operating Systems	Size	Mirror
Talend-DataPreparation-Free-Desktop-2.5.1.exe	2.5.1	June 2018	Main	Windows	185MB	US Europe
Talend-DataPreparation-Free-Desktop-2.5.1-apache.exe	2.5.1	June 2018	Main	Windows	110MB	US Europe
Talend-DataPreparation-Free-Desktop-2.5.1-apache.dmg	2.5.1	June 2018	Main	MAC	105MB	US Europe
Talend-DataPreparation-Free-Desktop-windows-2.5.1.zip	2.5.1	June 2018	Main	Windows	179MB	US Europe

Comment installer Talend Data Preparation sur Mac OS X?



Pour installer et démarrer la version Mac de Talend Data Preparation, veuillez suivre les étapes suivantes :

1

Double-cliquez sur le fichier **Talend-DataPreparation-Free-Desktop-2.5.dmg** pour ouvrir le dossier.

2

Glissez et déplacez l'icône Talend dans le dossier Applications.

3

Talend Data Preparation figure maintenant dans la liste de vos **Applications**. Ouvrez-le avec un double-clic.

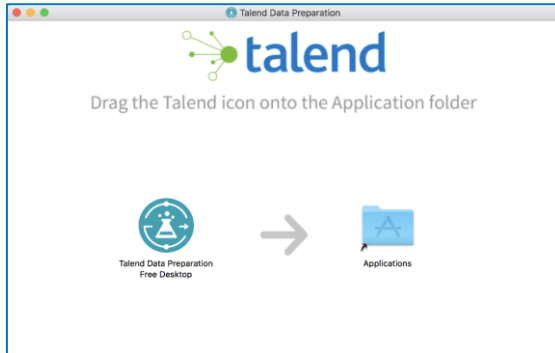
4

Pour désactiver **App Nap** et assurer des performances optimales, suivez cette procédure rapide :

1. Ouvrez le Terminal à partir du dossier
`/Applications/Utilit`
`ies.`

2. Exécutez la commande suivante :

```
defaults write  
org.talend.dataprep  
NSAppSleepDisabled -  
bool YES
```



Configuration de la langue de l'interface

Talend Data Preparation Free Desktop est intégralement traduit en français et japonais. La langue de l'application est l'anglais par défaut, mais une courte étape de configuration vous permet de choisir la langue de votre choix pour l'interface.

1

Ouvrez le fichier de configuration
<TDP_Installation_Path>
/dataprep/config/applic
ation.properties.

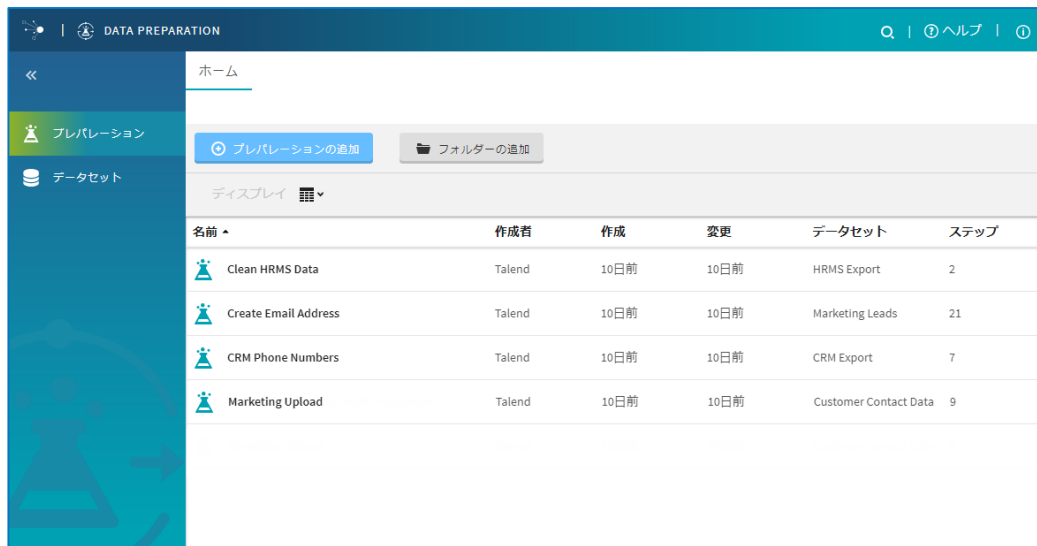
2

Pour le paramètre
dataprep.locale saisissez
l'une des trois valeurs supportées :

- **en-US** pour l'anglais
- **fr-FR** pour le français
- **ja-JP** pour le japonais

3

Redémarrez Talend Data
Preparation Free Desktop.



Page principale – Préparations et jeux de données

Une fois l'application démarrée, la première page qui apparaît à l'écran est la page "Preparations" :

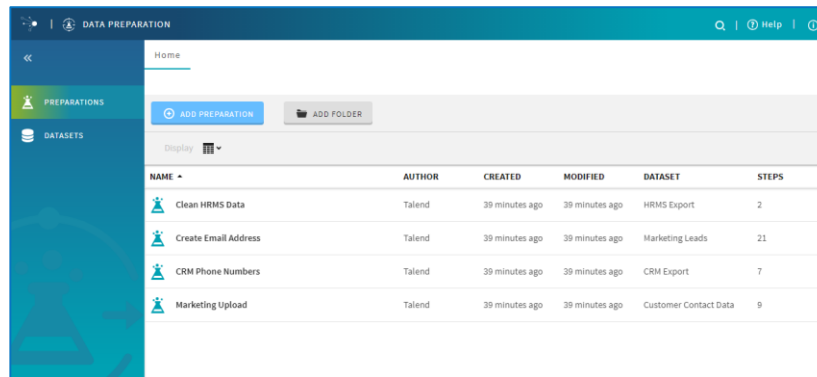
Dans la liste, vous verrez toutes les préparations sur lesquelles vous avez travaillé. Une préparation est le résultat des différentes étapes que vous avez appliqué pour nettoyer vos données. Vous pouvez exporter ce résultat en tant que fichier. Une préparation prend un jeu de données en entrée et applique la recette pour produire le résultat final. Les données d'origine ne sont jamais modifiées.

Depuis cette page, vous pouvez aussi accéder à la vue "Datasets" :

Vous verrez ici tous les jeux de données sur lesquels vous avez travaillé ou que vous avez importé. Les jeux de données peuvent être des fichiers en local ou à distance pouvant être importés dans Talend Data Preparation. Dans la version commerciale de Talend Data Preparation, ils peuvent également provenir d'une connexion à une base de données ou d'autres sources de données. Les jeux de données sont utilisés comme matériaux de base d'une ou plusieurs préparations.

À partir de cette page, vous pouvez :

- Ajouter de nouvelles préparations
- Organiser vos préparations en dossiers
- Importer et créer de nouveaux jeux de données
- Enregistrer vos jeux de données en tant que favoris

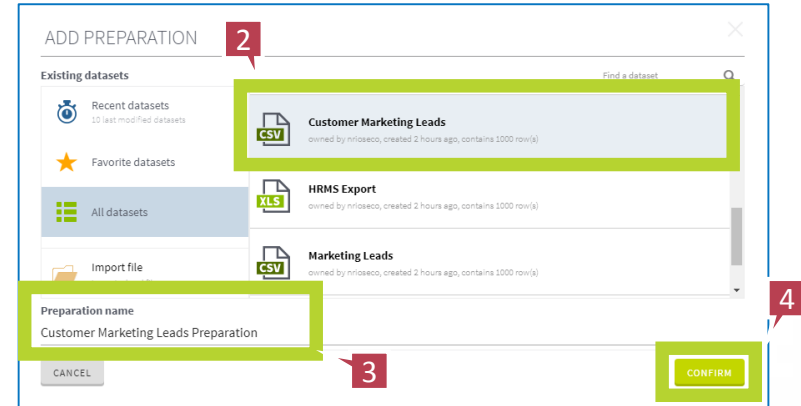
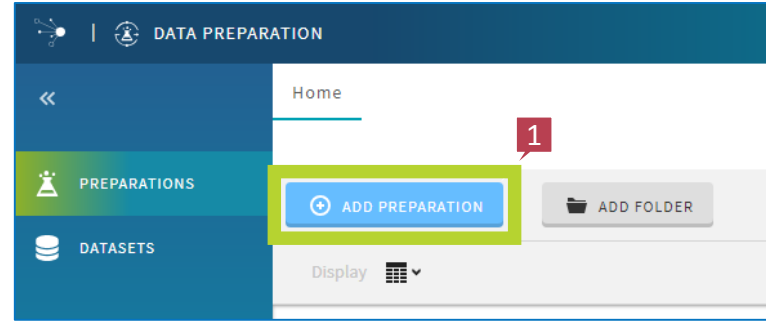


NAME	AUTHOR	CREATED	MODIFIED	DATASET	STEPS
Clean HRMS Data	Talend	39 minutes ago	39 minutes ago	HRMS Export	2
Create Email Address	Talend	39 minutes ago	39 minutes ago	Marketing Leads	21
CRM Phone Numbers	Talend	39 minutes ago	39 minutes ago	CRM Export	7
Marketing Upload	Talend	39 minutes ago	39 minutes ago	Customer Contact Data	9

Comment ajouter une préparation?

Pour commencer avec notre exemple:

1. Cliquez sur le bouton **Add Preparation** dans la vue **Preparations**.
2. La fenêtre **Add Preparation** s'ouvre. Cliquez sur le jeu de données **Customers Marketing Leads** dans la liste **All Datasets**.
3. Choisissez un **nom** pour votre préparation.
4. Cliquez sur **Confirm** pour ouvrir la préparation et commencer à nettoyer les données du jeu de données.



Visite guidée de Talend Data Preparation

Dans cette démo, nous vous montrerons des...

Exercices faciles
de nettoyage de
données

Manipulations
de données

Fonctions de
standardisation et
d'enrichissement
avancées

Exemples simples de nettoyage

Occupons nous d'abord de la colonne **Name**.

1. Cliquez sur l'en-tête de la colonne **Name**.
2. Tout en appuyant sur la touche **Ctrl**, cliquez sur l'en-tête de la colonne **last_name**. Les deux colonnes sont maintenant sélectionnées. Vous pouvez également utiliser **Shift + click** pour sélectionner plusieurs colonnes.
3. Le menu situé en haut à droite affiche la liste des **Fonctions** disponibles. Afin de corriger les données, vous pouvez choisir celle qui vous convient ou bien utiliser la fonction suggérée.
4. Selon la résolution de votre écran, vous devrez peut-être descendre dans la liste pour trouver la fonction **Change to upper case**. Passez votre souris sur la fonction pour afficher un aperçu des modifications. Cliquez sur la fonction pour appliquer les changements aux deux colonnes sélectionnées.

Dans cette étape nous nettoyons les champs contenant les prénoms des clients pour effectuer une standardisation de base. Ici, les prénoms commencent soit avec une minuscule, soit avec une majuscule. Les espaces redondants et les noms de famille sont reconnus comme formats incorrects.

The screenshot shows the Talend Data Preparation interface with a table of customer data. The columns are Name, First Name, Last Name, and Last Name. A context menu is open over the selected columns, listing various functions. Red arrows and numbers 1-4 indicate the sequence of actions:

1. Click on the header of the **Name** column.
2. While holding the **Ctrl** key, click on the header of the **last_name** column.
3. The context menu is open, showing a list of functions.
4. The mouse cursor is hovering over the **Change to upper case...** function.

The table data is as follows:

ID	Name	First Name	Last Name	Last Name	email	job_title	company	city	state	zip	country	date	campaign_id	lead_score
1	718143	Jason	Alexander	jalex	jalex@talend.com	Chemical Engineer	Talend	New York	NY	10001	USA	2011/2015	MOCKE_V1100L_lead	5
2	770396	Lillian	Simpson	lsimp	lsimp@talend.com	Desktop Support Tech	Canada	Victoria	BC	V8N 1K6	Canada	2/28/2015	BA_V1100L_lead	36
3	524952	WALTER	Ruiz	wruiz	wruiz@talend.com	Geological Engineer	Haitiers	Fairbanks	AK	99701	USA	3/18/2015	TRAIL_V1100L_lead	92
4	744980	Joshua	Hunt	jhunt	jhunt@talend.com	Financial Advisor	Dynex	Wilmington	DE	19804	USA	3/18/2015	MOCKE_V1100L_lead	79
5	404656	Mildred	Flores	mflor	mflor@talend.com	Nurse	Egipzian	Miami	FL	33139	USA	10/15/2015	MOCKE_V1100L_lead	46
6	9580018	Victor	Gonzalez	vgonz	vgonz@talend.com	State Associate	Mag	Charlotte	NC	28203	USA	10/12/2014	TRAIL_V1100L_lead	85
7	595042	Joshua	Simmons	jsimm	jsimm@talend.com	Structural Therapist	Dba	Jacksonville	FL	32202	USA	12/12/2015	TRAIL_V1100L_lead	49
8	149872	Beverly	Wright	bwrigh	bwrigh@talend.com	Biostatistical	Dynastec	Indianapolis	IN	46204	USA	8/19/2016 10:00:00	TRAIL_V1100L_lead	57
9	88928	Fred	Evans	fevans	fevans@talend.com	Director of Sales	Esac	Anchorage	AK	99501	USA	3/6/2015	MOCKE_V1100L_lead	24
10	78160	Joseph	Roberts	jroberts	jroberts@talend.com	Research Nurse	Gaboski	Las Vegas	NV	89102	USA	3/18/2015	MOCKE_V1100L_lead	77
11	3708	Steven	Johnson	sjohnson	sjohnson@talend.com	Senior Radiologist	Johnson	Waco	TX	76702	USA	12/29/2014	SA_V1100L_lead	7
12	95045	Wendy	Holliver	wholliver	wholliver@talend.com	Admission Specialist	Blumens	Bridgeport	CT	06610	USA	6/1/2015	SA_V1100L_lead	89
13	43871	Barbara	Goodman	bgoodman	bgoodman@talend.com	Admission Specialist	Shuffertag	Nacico	NC	27115	USA	MOCKE_V1100L_lead	46	
14	38838	Victor	Lee	vlee	vlee@talend.com	Librarian	Radlyns	Bees	OR	97103	USA	4/4/2015	TRAIL_V1100L_lead	18
15	68010	Catherine	Allison	callison	callison@talend.com	actuary	Rudolph	Houston	TX	77002	USA	11/2/2015	TRAIL_V1100L_lead	27
16	74270	Andrew	Wood	awood	awood@talend.com	Senior Sector	Fatzy	Los Angeles	CA	90001	USA	12/29/2014	MOCKE_V1100L_lead	78
17	61919	Samuel	Yam	syam	syam@talend.com	Structural Engineer	Genex	New York	NY	10001	USA	8/31/2015	MOCKE_V1100L_lead	46
18	21793	Bruce	Williams	brwilliams	brwilliams@talend.com	Help Desk Operator	Gaboski	Bridgeport	CT	06610	USA	6/4/2015	MOCKE_V1100L_lead	82
19														
20														
21														
22														
23														
24														
25														
26														
27														
28														
29														
30														

Exemples simples de nettoyage

Effectuer des opérations basiques de formatage et de nettoyage

Colonne Name (suite)

1. En observant les données, vous constaterez que des petits carrés blancs sont affichés avant ou après certains prénoms, comme par exemple "Joshua".
2. Pour supprimer ces carrés blancs, cherchez puis sélectionnez la fonction **Remove trailing and leading characters**.
3. Sans cocher la case **Create new column**, sélectionnez **Whitespace** dans la liste déroulante **Padding character** et cliquez sur **Submit**.

Certaines fonctions, comme celle-ci, vous offrent la possibilité d'appliquer la transformation dans une nouvelle colonne, en cochant la case **Create new column**. Si vous ne la cochez pas, la fonction s'appliquera dans la colonne sélectionnée.

The screenshot shows the Talend Data Preparation interface with a data table titled 'Customer Marketing Leads Preparation'. The table has columns for 'id', 'name', 'last_name', 'email', 'job_title', 'company', 'city', 'state', 'date', 'campaign_id', and 'lead_score'. A red arrow points from the 'name' column in the table to a zoomed-in view of the data. In this view, the name 'JOSHUA' is highlighted with a green box, and a red arrow points to a dropdown menu. The dropdown menu shows the function 'Remove trailing and leading characters...' selected, with a red arrow pointing to the 'SUBMIT' button. A third red arrow points to the 'Create new column' checkbox, which is unchecked. The 'Padding character' dropdown is set to 'Whitespace'.

id	name	last_name	email	job_title	company	city	state	date	campaign_id	lead_score
1	JOSHUA	HUNT	joshua.hunt@company.com	Software Engineer	Google	Mountain View	CA	2017-01-01	1001	5
2	MILDRED	SMITH	mildred.smith@company.com	Marketing Specialist	Microsoft	Redmond	WA	2017-02-01	1002	3
3	VICTOR	RODRIGUEZ	victor.rodriguez@company.com	Product Manager	Amazon	Seattle	WA	2017-03-01	1003	4
4	JOSHUA	SIMMONS	joshua.simmons@company.com	Business Development	Facebook	Foster City	CA	2017-04-01	1004	6
5	BEVERLY	WRIGHT	beverly.wright@company.com	Operations Manager	IBM	Armonk	NY	2017-05-01	1005	2
6	FRED	RODRIGUEZ	fred.rodriguez@company.com	Systems Administrator	Oracle	Redwood City	CA	2017-06-01	1006	4
7	JOSEPH	PETERSON	joseph.peterson@company.com	Quality Assurance	LinkedIn	Sunnyvale	CA	2017-07-01	1007	3

Recettes

1. Chaque fois que vous sélectionnez une fonction, elle s'ajoute automatiquement à la recette, située dans le panneau de gauche.
2. Pour supprimer un élément de la recette, placez le curseur sur la ligne correspondante et cliquez sur la corbeille.
3. Pour renommer une préparation, cliquez sur l'icône en forme de crayon et entrez un nouveau nom.
4. La recette peut être masquée en cliquant sur la flèche.
5. Pour exporter le résultat de votre préparation, cliquez sur **Export** et choisissez un type de fichier.

The screenshot shows the Talend Data Preparation interface. On the left, a recipe titled 'Customer Marketing Leads Preparation' is visible with three steps: 'Change to upper case on column Name', 'Change to upper case on column last_name', and 'Remove trailing and leading characters'. A red callout box '1' points to the pencil icon for renaming, '2' to the trash icon for removing a step, and '3' to the plus icon for adding a step. On the right, a data table is displayed with columns: id, Name (First Name, Last Name), email, job_title, company, city, state, and Airport. A red callout box '4' points to the eye icon for toggling the recipe, and '5' points to the 'EXPORT' button in the top right corner.

id	Name	last_name	email	job_title	company	city	state	Airport
	First Name	Last Name	Email					
1	KATHRYN	GARCIA	kgarcia14@g					
2	JASON	ALEXANDER	jalexander44@gmail.c	Chemical Engineer	Abata	Pearl City	HI	
3	LILLIAN	SIMPSON	lsimpson7@gmail.com	Desktop Support Tech	Camtmo	Wichita	KS	
4	WALTER	RUIZ	wruiz2@gmail.com	Geological Engineer	Yakitri	Fairbanks	AK	
5	JOSHUA	HUNT	jhuntmk@last_fm	Financial Advisor	Oyope	Wilmington	DE	
6	MILDRED	FLORES	mflores0@earthlink.	Nurse	Edgeblab	Miami	FL	
7	VICTOR	GONZALEZ	vgonzalez8@npr.org	Sales Associate	Ntag	Atlanta	GA	
8	JOSHUA	SIMMONS	jsimmons5@newyorker	Occupational Therapi	Oba	Jacksonville	FL	
9	BEVERLY	WRIGHT	bwright3@arizona.edu	Biostatistician	Skyoodle	Indianapolis	IN	
10	FRED	RODRIGUEZ	frodrigueznc@fotki.c	Director of Sales	Etodel	Anchorage	AK	
11	JOSEPH	PETERSON	jpeterson@sohu.com	Research Nurse	Gabcube	Las Vegas	NV	
12	DENISE	MARTIN	dmartin@java.com	Speech Pathologist	Zooncast	Nampa	ID	
13	JENNIFER	SULLIVAN	jsullivan4@lycos.co	Automation Specialis	Bluezoom	Bridgeport	CT	
14	RONALD	GONZALES	rgonzales5@apple.c	Automation Specialis	Shuffletag	Racine	WI	
15	VICTOR	COX	vcox9@virginia.edu	Librarian	Skalth	Bend	OR	

Comme vous avez créé cette préparation à l'aide du bouton **Add Preparation**, vous n'avez pas besoin de sauvegarder votre travail. Chaque nouvelle étape de préparation est automatiquement sauvegardée.

Vous pouvez créer et sauvegarder plusieurs préparations pour chaque jeu de données. N'oubliez pas que les données d'origine de votre jeu de données ne sont pas modifiées.

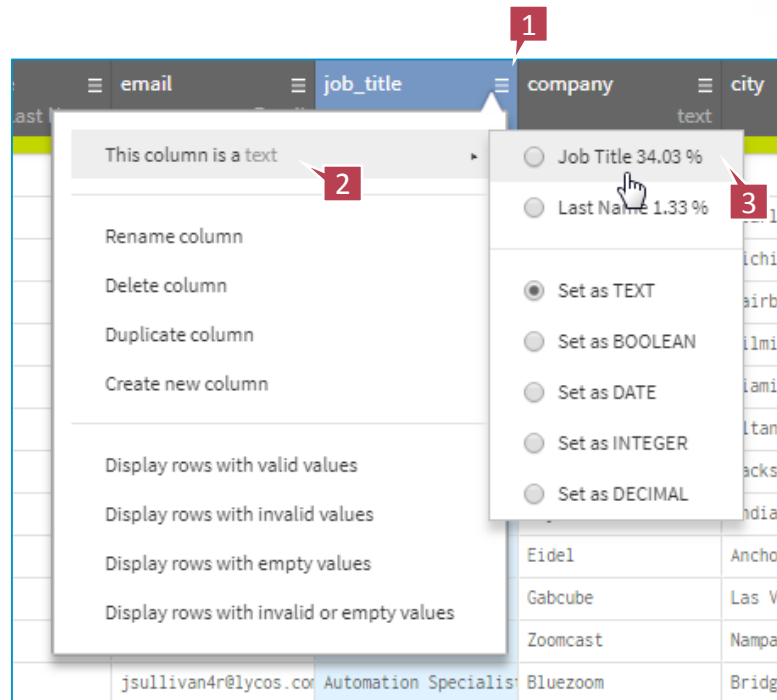
Type sémantique

Talend Data Preparation suggère automatiquement le type sémantique adapté pour les colonnes de vos jeux de données. Cela vous permettra de mieux identifier vos données. Vous pouvez toutefois modifier ces suggestions, en vous basant sur votre propre expérience.

Le type suggéré pour la colonne **job_title** est **Text**. Il faut donc le convertir en un type qui ait plus de sens, **Job Title** dans notre cas.

1. Dans l'en-tête de colonne, cliquez sur **l'icône de menu** pour sélectionner un nouveau type sémantique.
2. Passez votre souris au dessus de **This column is a text**.
3. Choisissez **Job Title**.

La version « Enterprise Edition » de Talend Data Preparation vous permet de créer des types sémantiques personnalisés. Elle vous permet également de modifier ou supprimer les types sémantiques par défaut.



Barre de qualité des données

En haut de chaque colonne, une barre mesure la qualité des données et indique par un code couleur le nombre de champs valides, vides ou invalides.

- **Vert** – Les données correspondent au type sémantique
- **Blanc** – Cellules vides
- **Orange** – Les données ne correspondent pas au type sémantique

id	Name	last_name	email	job_title
integer	First Name	Last Name	Email	text

Regardons de plus près la barre de qualité de la colonne **email**. En glissant le curseur sur chaque couleur, le nombre exact et le pourcentage des valeurs correspondant s'affichent.

- **Vert** – 979 cellules ont un format valide
- **Blanc** – 20 cellules sont vides
- **Orange** – 1 cellule a un format invalide

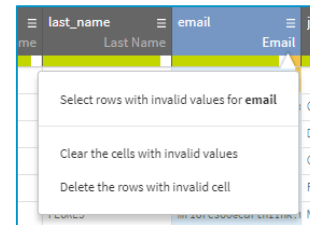
Pour sélectionner, supprimer ou vider les cellules ayant un format invalide, cliquez sur la barre colorée. Cliquez sur la section orange, puis sélectionnez **Select rows with invalid values** pour la colonne **email** afin de visualiser les adresses dont le format est invalide.

email	job_title
email	text
kgarcia14@g... jallexander44@gmail.com	Software Engineer

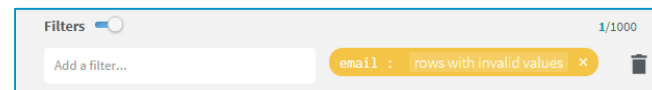
979 valid values (98%)

20 empty values (2%)

1 invalid values (0%)



N'oubliez pas de supprimer le filtre pour revenir à la liste complète.



Manipulation de texte basique

Pour filtrer les lignes invalides :

1. Cliquez sur l'en-tête de la colonne **state**.
2. En bas à droite de l'écran se trouve le graphique des modèles de données **Pattern**. En glissant le curseur sur chaque ligne, des analyses quantitatives vont apparaître. La ligne du haut indique que 911 enregistrements contiennent un code à deux lettres (correspondant à un État). **Vous pouvez cliquer sur l'une des barres pour filtrer ces enregistrements. Pour supprimer le filtre, cliquez sur le x dans le filtre, ou cliquez sur la poubelle dans la barre des filtres.**
3. Dans la barre de qualité, cliquez sur la **section orange**.
4. Cliquez sur **Select rows with invalid values for state**.
5. 7 lignes contenant des informations invalides s'affichent.

Ici nous avons nettoyé la valeur des champs ayant un format invalide. Vous apprendrez comment utiliser les graphiques pour filtrer les données mais aussi pour modifier des valeurs directement dans la grille des données.

The screenshot illustrates the process of filtering data in Talend Data Preparation. It shows a data grid with columns 'city', 'state', and 'date'. A red box labeled '1' highlights the 'state' column header. A second red box labeled '2' points to the 'Pattern' chart in the bottom right corner, which displays a bar chart with a value of 911 for the 'state' column. A third red box labeled '3' points to a dropdown menu that appears when clicking the 'state' header, with the option 'Select rows with invalid values for state' highlighted. A fourth red box labeled '4' points to the selected option in the dropdown menu. The data grid shows several rows with invalid state codes, such as 'AA', 'Aaaaa', 'aa', and 'A'.

city	state	date
Pearl City	us state code	11/2015
Wichita	KS	2/28/2015
Fairbanks	AK	7/15/2015
Wilmington	DE	3/16/2015
Miami	FL	10/15/2015
Atlanta	GA	17-12-2014
Jacksonville	FL	17-12-2015
Indianapolis	IN	01/01/2016 10:00:0
Eschorage		7/6/2015
		7/16/2015
		2/9/2014
		7/1/2015
		11.2015
		6.2015
		1/2/2015
		7/15/2
		3/16/2
		10/15/
		17-12-

Manipulation de texte basique

Filtrer et corriger les données :

1. Pour modifier le texte d'un champ, **double-cliquez sur une des cellules** contenant **Texas**. Changez **Texas** en **TX**. **N'appuyez PAS encore sur la touche Entrée !**
2. Sous la cellule que vous modifiez, cochez **Apply to all cell with this value** pour appliquer la modification aux autres cellules de même valeur. **Maintenant vous pouvez appuyer sur Entrée !** Vous venez de changer toutes les cellules avec la valeur **Texas** en **TX**.
3. Il reste deux lignes de données invalides. Regardez la liste des fonctions disponibles et **choisissez celle que vous souhaitez** utiliser pour corriger les codes invalides.
4. Une fois toutes ces actions effectuées, la **barre de qualité** de la colonne **state** sera uniquement **verte et blanche**.
5. Cliquez sur le **x** dans le filtre **state: rows with invalid values** pour revenir à la liste de données complète.

Filters: state : rows with invalid values x 7/1000

Add a filter...

	last_name	email	job_title	company	city	state	date	camp
	Last Name	Email	Job Title		text	Airport	US State Code	date
213	KENNEDY	ekennedy5@youtu.be	Executive Secretary	Flipopia	Cedar	TX	11/13/2015	HOCKEY
754	MATTHEWS	amatthewsg@soup.io	Staff Scientist	Eazzy	Dallas	TX		HOCKEY
756	LOPEZ	jlopezi0@geocities.jp	Clinical Specialist	Flipstorm	Austin			HOCKEY
757	CRAWFORD	ecrawfordjj@nasa.gov	Administrative Assis	Ozu	Dallas			BIKE_Y
765	SHAW	jshawpm@uiuc.edu	Occupational Therapi	Rooko	Piano			TRAIL
961	WEBB	rwebbrk@theguardian.	Administrative Assis	Thoughtmix	Dallas	Texas	10/28/2015	HOCKEY
985	WALKER					E	8/11/2015	TRAIL

state	date
us state code	
HI	22/11/
KS	2/28/2

Recettes

Chaque fonction utilisée a été ajoutée à la recette. La dernière étape nous dit que tous les champs ayant **Texas** comme État ont été changés en **TX**.

3 Remove trailing and leading characters on column Name

4 Change semantic domain on column job_title

5 Search and replace on column state

state : 1 rows with invalid values x

Create new column

Search for:

= Texas

Replace with:

TX

Overwrite entire cell

SUBMIT

6 Delete these filtered rows on column city

Manipulation numérique basique

Maintenant, passons à la colonne **lead_score**.

1. Sélectionnez la colonne **lead_score**. Il s'agit ici de champs de nombres entiers, mais **l'histogramme** à droite de l'écran nous dit que les données sont faussées par des valeurs plus grandes.
2. Cliquez sur la **barre bleue** tout à droite dans le graphique : 31 cellules ont la valeur 999. Il semblerait que la valeur par défaut soit réglée à 999. Cette fois, nous utiliserons la fonction **Fill cell with value**.
3. Tapez **Fill** dans la boîte de recherche en haut à droite de l'écran, puis sélectionnez **Fill cell with value**. Réglez la valeur à **0** et cliquez sur **Submit**.

Ici nous avons nettoyé et modifié des valeurs aberrantes dans un champ numérique. Vous apprendrez comment utiliser les graphiques pour filtrer les données, mais aussi pour modifier des valeurs directement dans la grille des données



The screenshot shows the Talend Data Preparation interface for the 'lead_score' column. The column is currently set to 'integer'. The 'Fill' tab is active, showing the 'Fill cells with value...' configuration. The 'Value' field is set to '0'. A 'SUBMIT' button is visible. Below the configuration, an 'HISTOGRAM' chart is displayed, showing the distribution of values. The x-axis ranges from -87 to 999, and the y-axis shows 'Occurrences' from 0 to 800. A red circle highlights the value '0' on the x-axis, with a red question mark next to it. A red box with the number '2' points to this area. Another red box with the number '3' points to the 'Fill cells with value...' configuration area. A red box with the number '1' points to the 'lead_score' column header in the data grid.

Manipulation numérique basique

Colonne lead_score, suite

1. En observant de plus près le graphique se rapportant à la colonne **lead_score**, vous remarquerez des valeurs négatives.
2. Puisqu'il est impossible d'avoir des scores négatifs, il faut les supprimer. Dans le menu des fonctions suggérées, sélectionnez **Calculate absolute value**. Cette fonction permet de garder la valeur des scores tout en éliminant le signe négatif.

The screenshot displays the Talend Data Preparation interface. At the top, a filter is applied to the 'lead_score' column: 'lead_score in [-87..0]'. Below this, a table shows data for various states and campaigns, with 'lead_score' values ranging from -40 to -87. A function menu is open, showing suggestions for 'lead_score'. The 'Calculate absolute value' option is highlighted with a green box and a red arrow labeled '2'. To the right, a histogram shows the distribution of 'lead_score' values, with a red arrow labeled '1' pointing to the negative values on the x-axis. The histogram's x-axis ranges from -87 to 105, and the y-axis shows 'Occurrences' from 0 to 160.

state	date	campaign_id	lead_score
HI	5/31/2015	HOOKEY_Y15Q02_wind	-40
AL	4/12/2015	TRAIL_Y14Q03_dots	-46
NM	4/11/2015	HOOKEY_Y15Q03_inns	-5
NC	4/3/2015	TRAIL_Y15Q03_c1ad	-41
UT	4/23/2015	SKI_Y15Q03_fe11	-65
CA	4/1/2015	TRAIL_Y15Q03_nope	-42
CA	7/9/2015	SKI_Y15Q03_laws	-87

Nettoyage et formatage de dates

Occupons nous maintenant de la colonne **date**.

1. Cliquez sur l'en-tête de la colonne **date**, puis sur **Pattern** à droite de l'écran. Vous pourrez visualiser aisément tous les différents formats de date et de masquage utilisés. Certaines dates sont au format européen, d'autres au format américain, certaines contiennent des slashes, d'autres des tirets.
2. Pour standardiser les dates, cliquez sur **Change date format** dans la liste des fonctions suggérées. Sélectionnez un format parmi ceux proposées ou insérez celui de votre choix, puis cliquez sur **Submit**.

Combien de fois sommes-nous confrontés à des tableaux comportant toute sorte de formats et standards de dates extravagants ? Nous savons tous qu'Excel peut reformater les champs date, mais quand ils présentent un mélange de formats et différents masquages, Excel ne gère plus !

The screenshot displays the Talend Data Preparation interface. A table with columns 'state', 'date', 'campaign_id', and 'lead_score' is visible. The 'date' column header is highlighted with a green box labeled '1'. To the right, a 'date' column configuration panel shows a list of 'SUGGESTIONS' with 'Change date format...' highlighted by a green box labeled '2'. Below this, a 'Change date format...' dialog box is open, showing the 'Current format' as 'I don't know, best guess' and the 'New format' as 'Other' with 'MM.dd/yyyy' selected. A green box labeled '3' highlights the 'Submit' button in the dialog box.

Nettoyage et formatage de dates

Modifier la recette, c'est facile.

1. Dans la **recette** à gauche, cliquez sur la dernière action.
2. Dans la liste déroulante de la fonction **Change date format**, choisissez **Other** pour saisir un format personnalisé. Saisissez **dd-MMMM-yyyy** (attention aux minuscules et aux majuscules, cette fonction y est sensible).
3. Cliquez sur **Submit**, les modifications seront alors appliquées. Vous pouvez supprimer une étape de la recette (cliquez sur la corbeille) ou la désactiver (cliquez sur la coche verte).
4. Vous pouvez également **réordonner les étapes de votre recette par glisser-déposer**. Cela vous fera gagner du temps si, par exemple, vous réalisez qu'une colonne sur laquelle vous avez appliqué une fonction contenait encore des données invalides à nettoyer.

The screenshot shows the Talend Data Preparation interface. On the left, the 'Recipe' pane displays two actions: '8 Calculate absolute value on column lead_score' and '9 Change date format on column date'. The 'Change date format' action is selected, and its configuration is shown in a modal window. The 'Current format' is 'I don't know, best guess' and the 'New format' is 'Other'. A green box highlights the 'Your format:' field containing 'MM.dd.yyyy' and the 'SUBMIT' button. A red '2' is next to this box. A second green box highlights the 'Your format:' field containing 'dd-MMMM-yyyy' and the 'SUBMIT' button. A red '3' is next to this box. On the right, the 'Filters' pane shows a table with columns 'email', 'job_title', and 'company'. The table contains 12 rows of data. A red '1' is next to the 'Change date format' action in the recipe pane.

	email	job_title	company
2	jalexander44@gmail.c	Chemical Engineer	Abata
3	lsimpsonf7@gmail.com	Desktop Support Tech	Camibo
4	wruiz1z@gmail.com	Geological Engineer	Yakitri
5	jhuntmk@last.fm	Financial Advisor	Oyope
6	mfloreso6@earthlink.	Nurse	Edgeblab
7	vgonzalez8c@npr.org	Sales Associate	Ntag
8	jsimmons5@newyorker	Occupational Therapi	Oba
9	bwright3@arizona.ed	Biostatistician	Skynoodle
10	frodrigueznc@fotki.c	Director of Sales	Eidel
11	jpeterosnm@sohu.com	Research Nurse	Gabcube
12	dmartint@java.com	Speech Pathologist	Zoomcast
12	teullian4d@lucor.co	Automation Specialis	Bluezoom
			Shuffletag
			Skalith
			Rhyloo
			Teev

Masquage de données

Vous pouvez facilement masquer des données sensibles.

1. Cliquez sur l'en-tête de la colonne **email** pour sélectionner son contenu.
2. Dans le liste des fonctions, recherchez **Mask data (Obfuscation)**.
3. Cliquez dessus pour appliquer la fonction sur les adresses email.
4. Tous les caractères avant @ sont remplacés par XXX, alors que le reste ne change pas. C'est l'effet de la fonction de masquage de données sur les cellules dont le type sémantique est l'email. Mais l'effet du masquage de données sera différent selon le type sémantique de la colonne.

Lorsque vous manipulez des données sensibles telles que des noms, des adresses, des numéros de carte de crédit ou de sécurité sociale, vous pouvez avoir besoin de masquer ces données. Pour protéger les données d'origine, utilisez la fonction de masquage de données afin de générer des alternatives fonctionnelles.

The screenshot illustrates the process of applying the 'Mask data (Obfuscation)' function to an 'email' column in Talend Data Preparation. The interface shows a data table with columns like 'id', 'name', 'last_name', 'email', 'job', 'company', 'city', 'state', 'date', 'campaigns_id', and 'lead_source'. The 'email' column header is highlighted with a red box and labeled '1'. A function list on the right is open, showing 'Mask data (obfuscation)' selected, with a red box and label '2'. A preview window shows the 'email' column with masked values like 'XXXXXXXXXX' and 'XXXXXXXXXX@gmail.com', with a red box and label '3'. A red arrow points from the function list to the preview window. A red box and label '4' highlights the 'email' column header in the preview window.

Le Data blending

Le Data Blending consiste à combiner les données issues de différentes sources. Cette fonction vous permet d'importer des données d'un jeu de données et de les ajouter à celui sur lequel vous êtes en train de travailler.

1. Cliquez sur l'icône de **Data Blending**.
2. La liste comprenant les jeux de données que vous avez enregistrés et d'autres fichiers préchargés est disponible en cliquant sur l'icône **+**.
3. Cochez **Business Unit Regions With States** et cliquez sur **Add**.



The screenshot shows the 'Select dataset(s) to add to lookup' dialog box in the Talend Data Preparation interface. The dialog lists several datasets, with 'Business Unit Regions With States' selected. Red callout boxes 1, 2, 3, and 3b indicate the steps described in the text.

Dataset Name	Selected
Customers	<input type="checkbox"/>
States	<input type="checkbox"/>
Emails Reference	<input type="checkbox"/>
CRM Export	<input type="checkbox"/>
Business Unit Regions With States	<input checked="" type="checkbox"/>
HMES Export	<input type="checkbox"/>
Marketing Leads	<input type="checkbox"/>

Le Data blending

Data blending, suite

1. Cliquez sur la **colonne que vous souhaitez importer et combiner**, dans ce cas, la colonne **state** dans le jeu de données en cours d'utilisation.
2. Ajoutez les régions en cliquant sur **Add to dataset** au-dessous de l'en-tête de la colonne **Region**.
3. **Passez votre curseur sur le bouton Confirm** afin de prévisualiser les modifications qui seront affichées en vert. Pour appliquer ces modifications, cliquez sur **Confirm**.

The screenshot shows the Talend Data Preparation interface for a dataset named 'Customer Marketing Leads Preparation'. The main data table has columns: id, Name (first_name, last_name), email, job_title, company, city, Airport, state, US State Code, Region, date, and campaign_id. A 'Filters' section is at the top left. Below the table, a 'Region' dropdown menu is open, showing 'US State Codes' and a list of regions (AK, HI, VT, NH, ME, CT, RI, MA, NY, NJ, PA, OH, WV, KY, TN, MS, AL, GA, SC, NC, VA, MD, DE, DC, WV, VA, NC, SC, GA, FL, HI, AK). A 'Confirm' button is highlighted in the bottom right corner. Red callout boxes with numbers 1, 2, and 3b point to the 'state' column, the 'Add to Dataset' button, and the 'Confirm' button respectively.

Regrouper et standardiser

La fonction pour regrouper et standardiser les données vous permet de localiser des cellules ayant un contenu texte similaire afin de les réunir et d'harmoniser le texte.

1. Cliquez sur l'en-tête de la colonne **job_title**.
2. Le graphique en bas à droite montre la grande quantité de dénominations de poste similaires dans le fichier. Afin d'harmoniser, nous devons tout d'abord regrouper les dénominations similaires.
3. Dans la barre de recherche, saisissez **group**.
4. Cliquez sur la fonction **Find and group similar text**.

The screenshot displays the Talend Data Preparation interface. On the left, a table with columns 'job_title', 'company', 'city', and 'state' is shown. The 'job_title' column is selected, indicated by a red '1'. On the right, the 'job_title' column is expanded, showing a search bar with 'group' entered (red '3') and the 'Find and group similar text...' function selected (red '4'). Below this, a bar chart (red '2') shows the frequency of various job titles. The chart has a 'ROW COUNT' dropdown and a scale from 0 to 20. The most frequent job title is '(Empty)' with a count of approximately 20. Other frequent titles include 'Occupational Therapist', 'Database Administrator', 'VP Marketing', 'Financial Advisor', 'Web Designer', 'Software Consultant', 'Human Resources Manager', 'Librarian', 'Senior Financial Analyst', 'Geological Engineer', 'Structural Engineer', 'VP Sales', 'VP Product Management', and 'Environmental Tech'.

job_title	company	city	state
Chemical Engineer	Abata	Pearl City	HI
Desktop Support Tech	Camimbo	Wichita	KS
Geological Engineer	Yakitri	Fairbanks	AK
Financial Advisor	Oyope	Wilmington	DE
Nurse	Edgeblab	Miami	FL
Sales Associate	Ntag	Atlanta	GA
Occupational Therapi	Oba	Jacksonville	FL
Biostatistician	Skynoodle	Indianapolis	IN
Director of Sales	Eidel	Anchorage	AK
Research Nurse	Gabcube	Las Vegas	NV
Speech Pathologist	Zooncast	Nampa	ID
Automation Specialis	Bluezoom	Bridgeport	CT
Automation Specialis	Shuffletag	Racine	WI
Librarian	Skalith	Bend	OR
Actuary	Rhylou	Manhattan	NY
Senior Editor	Tazzy	Columbus	GA
Structural Engineer	Dynava	Overland Park	KS
Help Desk Operator	Gabtune	Orange	CT
Senior Sales Associa	Npath	Cheshire	CT
VP Marketing	Oozz	New Haven	CT
Research Associate	Tavu	Prospect	CT
Tax Accountant	Devbug	New Haven	CT
Professor	Blogpad	East Lyme	CT
Financial Analyst	Chatterpoint	New Haven	CT
Systems Administrato	Fivebridge	Greenville	DE
Junior Executive	Kwinbee	Wilmington	DE
Librarian	Feedfish	Pike Creek	DE
Nurse	Youfeed	Greenville	DE

Trouver et regrouper des données texte similaires

Regrouper et standardiser, suite

1. Toutes les dénominations de poste similaires sont regroupées dans la deuxième colonne.
2. La troisième colonne suggère la dénomination de poste qui pourrait **remplacer** celles de la deuxième colonne. Vous pouvez utiliser **le menu déroulant pour choisir une autre dénomination ou bien insérer celle de votre choix**.
3. Si vous ne souhaitez pas modifier une dénomination spécifique, **décochez** la case devant la dénomination en question.
4. Si vous ne souhaitez pas modifier un groupe de dénominations de poste, **décochez** la case dans la première colonne.
5. Cliquez sur **Submit** quand vous avez terminé.

FIND AND GROUP SIMILAR TEXT

Replace all similar values with the right one (i.e. cluster on fuzzy matching)

<input checked="" type="checkbox"/>	These values have been found 1	This value will be kept
<input checked="" type="checkbox"/> 3	<input checked="" type="checkbox"/> Health Coach <input checked="" type="checkbox"/> Health Coach	Replace value: Health Coach
<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/> Administrative Assistant <input checked="" type="checkbox"/> Administrative Officer	Replace value: Administrative Assistant 2
<input checked="" type="checkbox"/> 4	<input checked="" type="checkbox"/> Account Executive <input checked="" type="checkbox"/> Account Representative <input checked="" type="checkbox"/> Account Representative <input checked="" type="checkbox"/> Accountant <input checked="" type="checkbox"/> Accounting Assistant	Replace value: Accountant

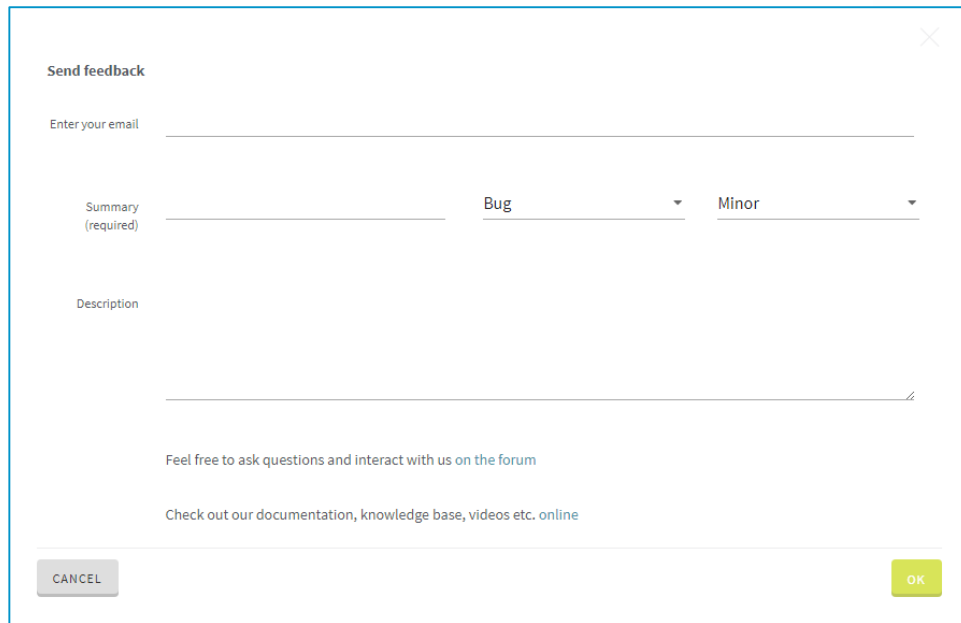
5 SUBMIT

Commentaires

Vos commentaires sont très importants pour nous.

Nous souhaitons savoir si nos services répondent à vos besoins et si nous vous les proposons de manière efficace.

Pour laisser un commentaire, cliquez sur l'icône **Information** et sélectionnez **Feedback** dans le menu déroulant. Complétez le formulaire avec votre adresse email et vos observations. N'hésitez pas à utiliser également les liens vers le forum et la base de connaissances présents dans le formulaire.



The screenshot shows a 'Send feedback' dialog box with a close button (X) in the top right corner. The form contains the following fields and elements:

- Send feedback** (Title)
- Enter your email** (Text input field)
- Summary (required)** (Text input field)
- Bug** (Dropdown menu)
- Minor** (Dropdown menu)
- Description** (Text area)
- Feel free to ask questions and interact with us on the forum** (Text)
- Check out our documentation, knowledge base, videos etc. online** (Text)
- CANCEL** (Button)
- OK** (Button)

Conclusion



Maintenant, vos données sont prêtes :

- **Pour l'analyse.** Vos données sont nettoyées et standardisées, et vous avez exporté les résultats de votre préparation. Vous pouvez par exemple analyser le potentiel de vos prospects par date ou par région dans Excel ou dans Tableau.
- **Pour l'intégration.** Les données sont nettoyées et formatées, vous pouvez les charger dans une application de CRM ou d'automatisation du marketing, telle que Marketo ou Salesforce.

La bonne nouvelle, c'est que...

Avec Talend, les données sont à un clic de vos tâches quotidiennes.

Et après?

Maintenant que vous avez appris à ajouter vos jeux de données dans l'application, à effectuer vos propres préparations... vous pouvez transformer vos tâches quotidiennes en activités guidées par les données.



